CSE 158 – Lecture 11

Web Mining and Recommender Systems

Text mining Part 2

Recap: Prediction tasks involving text

What kind of quantities can we model, and what kind of prediction tasks can we solve using **text?**

Prediction tasks involving text

Does this article have a positive or negative sentiment about the subject being discussed?

What can stop US Postal Service trucks? The inexorable march of time

The ageing fleet of delivery vehicles is long past due an overhaul. Among the common-sense upgrades employees want: air conditioning and more workspace

Neither snow nor rain nor heat nor gloom of night stays these trucks - but time, it turns out, will. Photograph: Bill Sikes/AP

For the better part of the last 30 years, the flatulent buzz of the US Postal Service's boxy delivery vans - audible as they lighted from mailbox to mailbox - has been a familiar sound to most Americans. Neither snow nor rain nor heat nor gloom of night stays the USPS's mail trucks from the swift completion of their appointed

Prediction tasks involving text

What is the category/subject/topic of this article?

Apple Is Forming an Auto Team

By BRIAN X. CHEN and MIKE ISAAC FEB. 19, 2015

Email

Share

Tweet

Save

More

SAN FRANCISCO — While Apple has been preparing to release its first wearable computers, the company has also been busy assembling a team to work on an automobile.

The company has collected about 200 people over the last few years — both from inside Apple and potential competitors like Tesla — to develop technologies for an electric car, according to two people with knowledge of the company's plans, who asked not to be named because the plans were private.

The car project is still in its prototype phase, one person said, meaning it is probably many years away from being a viable product and might never reach the mass market if the quality of the vehicle fails to impress Apple's executives.

Electric car batteries being prepared for shipment at the A123 Systems plant in Livonia, Mich in 2012. Apple has hired engineers from A123 Systems. Stephen McGee for The New York Times

It could also go nowhere if Apple struggles to find a compelling business opportunity in automobiles, a business that typically has much lower sales margins than

Prediction tasks involving text

Which of these reviews am I most likely to agree with or find helpful?

Most Helpful Customer Reviews

1,900 of 1,928 people found the following review helpful

Le Creuset on a budget

By N. Lafond on October 24, 2007

Color Name: Caribbean Blue | Size Name: 6 qt | **Verified Purchase**

Enamel on cast iron cookware like this, was, until recently, only available from makers like Le Creuset. Lately, several lower cost makers have come on the scene, like Target and Innova. The new budget priced Lodge cookware is in the same price range as the low cost alternatives but completely out performs them.

I have all of the brands I have mentioned. The Lodge is the same weight as the Le Creuset which is much heavier than the other budget models. The ridge where the lid and sides meet is a matt black porcelain on the Lodge and Le Creuset but is just exposed cast iron for the other budget models (which leads to rusting if you are not careful). The porcelain resists staining (even tomato sauces) in the Lodge and Le Creuset but the other budget models stain very easily. And finally, the Lodge and Le Creuset maintain a very polished interior finish that resists sticking which others do not. So, I see no performance differences at all between the Le Creuset and the Lodge whereas the comparably priced budget models are certainly inferior.

If you plan of using these pots very heavily (every day for example) you might want to upgrade to the higher priced Lodge product. It has 4 coatings of enamel as opposed to 2 in this model. But if you use them once or twice a week I dont think you will need the added wear resistance.

47 Comments | Was this review helpful to you? Yes No

1,105 of 1,164 people found the following review helpful

OK pot, Great Price. Some flaws.

By J. G. Pavlovich on March 2, 2008

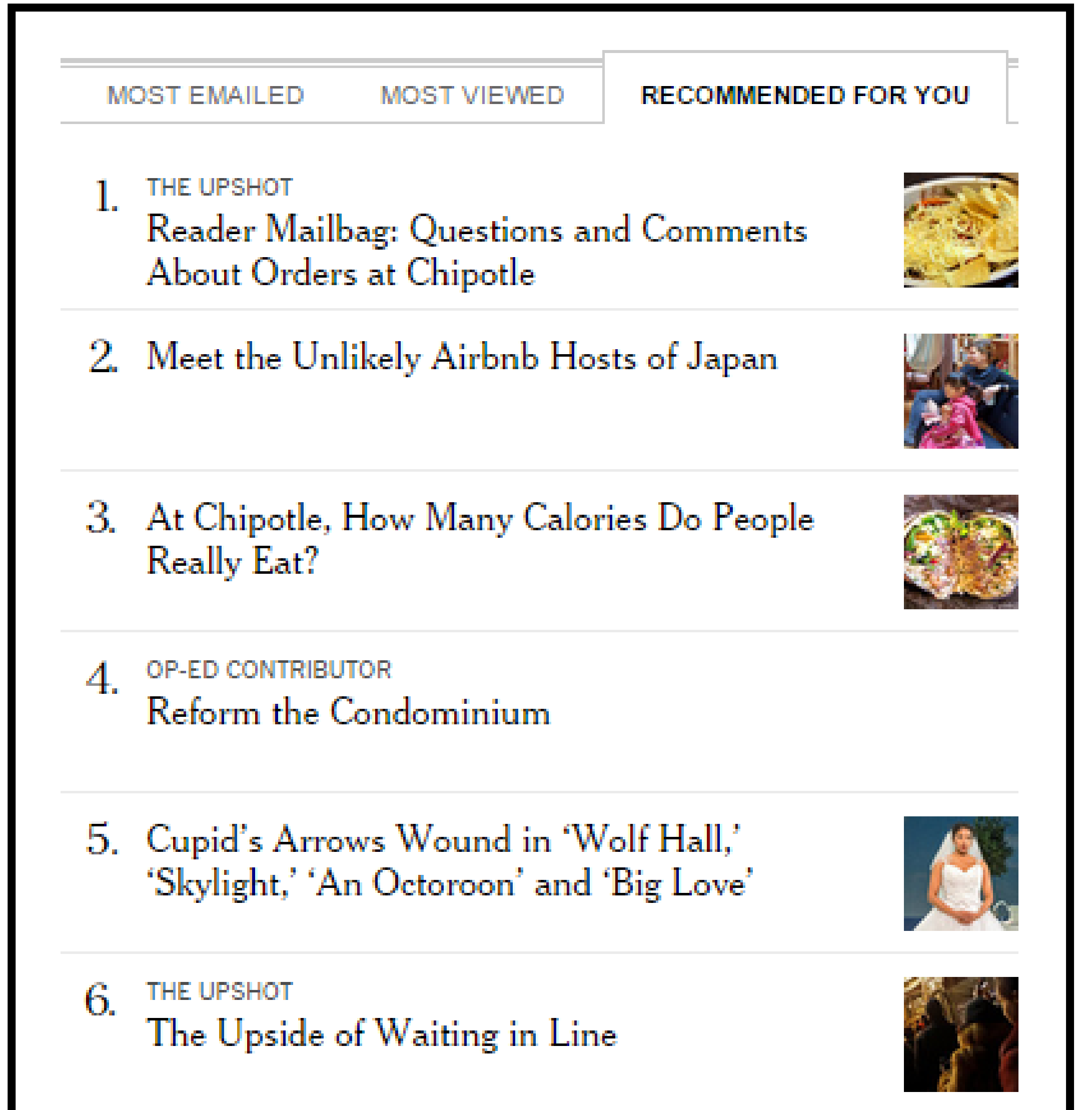
Color Name: Island Spice Red | Size Name: 6 qt | **Verified Purchase**

This is a terrific value. The quality and performance match my Le Creuset pieces at a fraction of the price. The only slight design flaw I have found is that the rounded bottom makes browning large pieces of meat awkward. Other than that I have no complaints. Even heating. Easy clean up. I use it several times a week.
UPDATE: I found a second minor problem. The inside rim of the lid has a couple of raised spots which prevent the lid from seating tightly. This causes steam to escape much faster than I would like during a long braise or stew.

Update 2: Three years in I am dropping my rating to three stars. It's still a decent pot at a bargain price, but it will not be an heirloom piece like my Le Creuset. The loose fitting lid turns

Prediction tasks involving text

Which of these articles are relevant to my interests?



Prediction tasks involving text

Find me articles similar to this one

Meatloaf That Conquers the Mundane

FEB. 13, 2015

City Kitchen

By DAVID TANIS

Email

Share

Tweet

Pin

Save

More

I was raised on Midwestern meatloaf. My mother's dependable recipe did not vary: Ground beef, grated onion and carrot and a little oatmeal were the main ingredients, along with a dash of "seasoned salt." A ribbon of bottled chili sauce ran down a gully in the center.

Served hot, accompanied by Tater Tots, it was dinner. Served cold for lunch, it was always a sandwich on white bread, with potato chips on the side. It was usually moist and tasty but never remarkable, and there was no way you could call it anything but meatloaf.

Do I harbor a kind of nostalgia for it? Yes. But would I use that recipe now? I think not.

Evan Sung for The New York Times

I have a friend from Brussels who loves to entertain. Of his dinner party repertoire, one dish is most requested and admired. It is pain de veau, served with a vermouth-splashed mushroom sauce. In French, it sounds elegant. Translated into English — veal loaf — it sounds dull.

The Italian word for meatloaf is polpettone. (Polpette are Italian meatballs; polpettine are meatballs, too, but more diminutive.) This substantial family-size meatball, whether ovoid or elongated, plain or fancy, served with tomato sauce or not, is beloved both in Italy and in Italian communities throughout the world. Aside from its melodic, polysyllabic name, polpettone is always well seasoned, prepared with care and served with gusto.

It is usually a combination of different kinds of ground meat, typically beef, pork and veal in equal parts. Grated cheese and herbs are

RELATED COVERAGE

City Kitchen: How to Make Polpettone, Step by Step FEB. 13, 2015

RECIPES FROM COOKING

Polpettone with Spinach and Provolone

By David Tanis

related articles

Feature vectors from text

Bag-of-Words models

The Peculiar Genius of Bjork

CULTURE | BY EMILY WITT | JANUARY 23, 2015 11:30 AM

Solo musician or master collaborator? For her new album, Bjork has merged the two sides of her artistry to create a new experience of music — again.

F_text = [150, 0, 0, 0, 0, 0, … , 0]

a

aardvark

zoetrope

musician, who creates her music in an emotional cocoon, tinkering with technologies, concepts and feelings; and Bjork the producer and curator, who seeks out

Feature vectors from text

Bag-of-Words models

Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.	yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma Medium tan

These two documents have **exactly** the same representation in this model, i.e., we're completely **ignoring** syntax.
This is called a "bag-of-words" model.

Feature vectors from text

Find the most common words...

```
counts = [(wordCount[w], w) for w in wordCount]
counts.sort()
counts.reverse()

words = [x[1] for x in counts[:1000]]
```

Feature vectors from text

And do some inference!

e.g.: Sentiment analysis

Let's build a predictor of the form:

$$f(\text{text}) \rightarrow \text{rating}$$

using a model based on linear regression:

$$\text{rating} \simeq \alpha + \sum_{w \in \text{text}} \text{count}(w) \cdot \theta_w$$

Code: http://jmcauley.ucsd.edu/cse258/code/week5.py

CSE 158 – Lecture 11

Web Mining and Recommender Systems

TF-IDF

Distances and dimensionality reduction

When we studied recommender systems, we looked at:

- Approaches based on measuring similarity (cosine, jaccard, etc.)
- Approaches based on dimensionality reduction

Today we'll look at the same two concepts, but using textual representations

Finding relevant terms

So far we've dealt with huge vocabularies just by identifying the **most frequently occurring** words

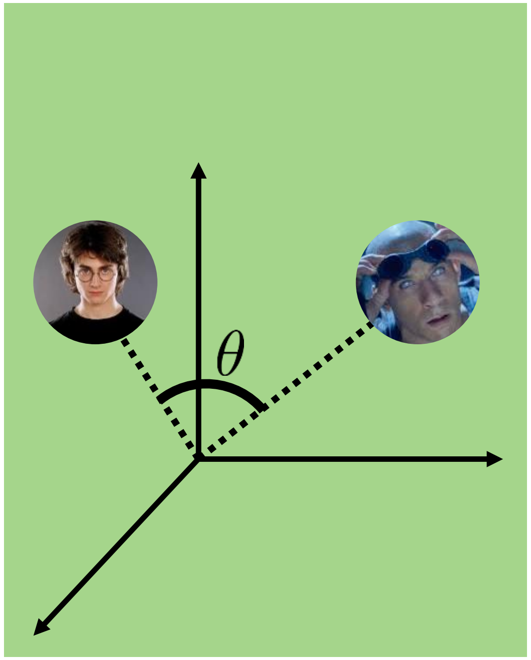
But! The most informative words may be those that occur very rarely, e.g.:

- Proper nouns (e.g. people's names) may predict the content of an article even though they show up rarely
- Extremely superlative (or extremely negative) language may appear rarely but be very predictive

Finding relevant terms

e.g. imagine applying something like cosine similarity to the document representations we've seen so far

e.g. are (the features of the reviews/IMDB descriptions of) these two documents "similar", i.e., do they have high cosine similarity



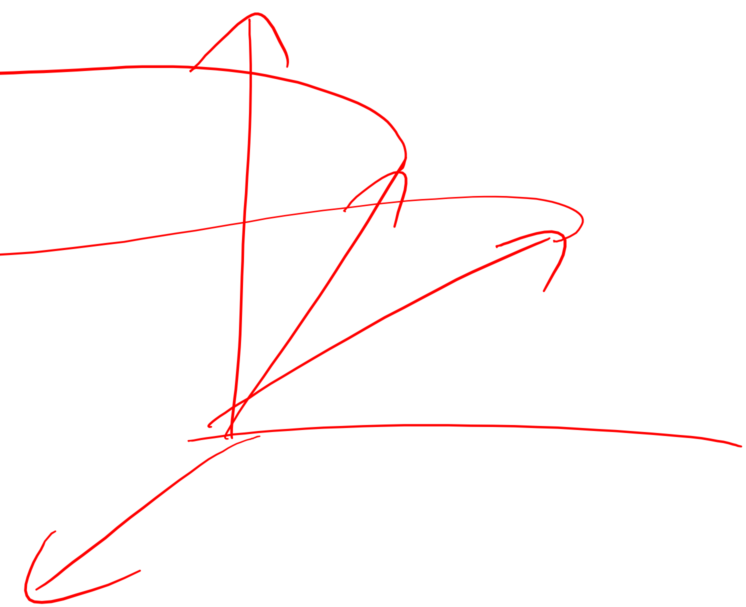
Finding relevant terms

e.g. imagine applying something like cosine similarity to the document representations we've seen so far

[652 458 ... 2 1]

[352 721 0 1]

a of the .... action - scifi



Finding relevant terms

So how can we estimate the “relevance” of a word in a document?

e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn’t ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we’ve got all the scoop on why Katy is set in her ways. Will these two *ever* bury the hatchet? Katy Perry & Taylor Swift Still Fighting? “Taylor’s tried to reach out to make amends with Katy, but Katy is not going to accept it nor is she interested in having a friendship with Taylor,” a source tells HollywoodLife.com exclusively. “She wants nothing to do with Taylor. In Katy’s mind, Taylor shouldn’t even attempt to make a friendship happen. That ship has sailed.” While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor’s managed to mend the fences with Katy’s BFF Diplo, but it looks like Taylor and Katy won’t be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That’s not too much to ask

Finding relevant terms

So how can we estimate the "relevance" of a word in a document?

e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn't ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we've got all the scoop on why Katy ill these two *ever* bury the hatchet? Katy Perry & Taylor Swift Still Fighting? "Tay t to make amends with Katy, but Katy is not going to accept it nor is she inte ndship with Taylor," a source tells HollywoodLife.com exclusively. "She wan Taylor. In Katy's mind, Taylor shouldn't even attempt to make a friendship happen. That ship has sailed." While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor's managed to mend the fences with Katy's BFF Diplo, but it looks like Taylor and Katy won't be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That's not too much to ask

"the" appears 12 times in the document

Finding relevant terms

So how can we estimate the “relevance” of a word in a document?

e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn’t ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we’ve got all the scoop on why
Katy … ill t … e hatchet? Katy Perry & Taylor Swift Still Fighting?
“Tay … ut to … Katy, but Katy is not going to accept it nor is she
inte … ends … urce tells HollywoodLife.com exclusively. “She
wan … Tay … ylor shouldn’t even attempt to make a friendship
happen. That ship has sailed.” While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor’s managed to mend the fences with Katy’s BFF Diplo, but it looks like Taylor and Katy won’t be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That’s not too much to ask

“the” appears 12 times in the document

“Taylor Swift” appears 3 times in the document

Finding relevant terms

So how can we estimate the “relevance” of a word in a document?

Q: The document discusses “the” more than it discusses “Taylor Swift”, so how might we come to the conclusion that “Taylor Swift” is the more relevant expression?

A: It discusses “the” **no more** than other documents do, but it discusses “Taylor Swift” **much more**

Finding relevant terms

Term frequency & document frequency

Term frequency ~ How much does the term appear in the document

Inverse document frequency ~ How "rare" is this term across all documents

Finding relevant terms

Term frequency & document frequency

$tf(w, d)$ = how many times w appears in d

word

doc

$df(w, D)$ = how many ~~times~~ ~~w across all $d \in D$~~ docs $d \in D$ contain w

set of docs

Term frequency & document frequency

"Term frequency": $tf(t,d)$ = number of times the term *t* appears in the document *d*

e.g. tf("Taylor Swift", that news article) = 3

"Inverse document frequency": $idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|}$

term (e.g. "Taylor Swift")   set of documents

"Justification": $P(t|D) = \frac{|\{d \in D : t \in d\}|}{N}$ so $idf(t,D) = -\log P(t|D)$

Finding relevant terms

Term frequency & document frequency

TF-IDF is high → this word appears much more frequently in this document compared to other documents

TF-IDF is low → this word appears infrequently in this document, or it appears in many documents

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Term frequency & document frequency

tf is sometimes defined differently, e.g.:

$$tf'(t,d) = \delta(t \in d)$$

$$tf''(t,d) = \frac{\text{frequency of word}}{\text{frequency of most common word in document}}$$

Both of these representations are invariant to the document length, compared to the regular definition which assigns higher weights to longer documents

Finding relevant terms

How to use TF-IDF

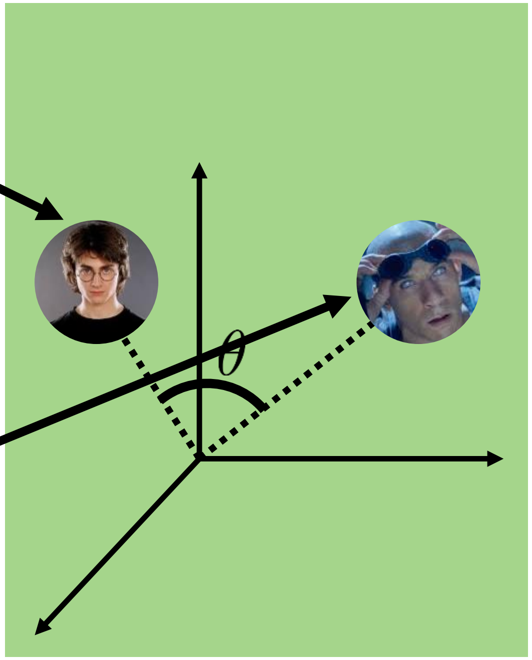
“fantasy”

[0,0,0.01,0,0.6,…,0.04,0,3,0,159.1,0]

“the” “and”

[180.2,0,0.01,0.5,0,…,0.02,0,0.2,0,0,0]

“action”



- Frequently occurring words have little impact on the similarity
- The similarity is now determined by the words that are most “characteristic” of the document

Finding relevant terms

But what about when we're **weighting** the parameters anyway?

e.g. is:

$$\text{rating} \simeq \alpha + \sum_{w \in \text{text}} \text{count}(w) \cdot \theta_w$$

really any different from:

$$\text{rating} \simeq \alpha + \sum_{w \in \text{text}} \mathit{tfidf}\,(w, d, D) \cdot \theta_w$$

after we fit parameters?

Finding relevant terms

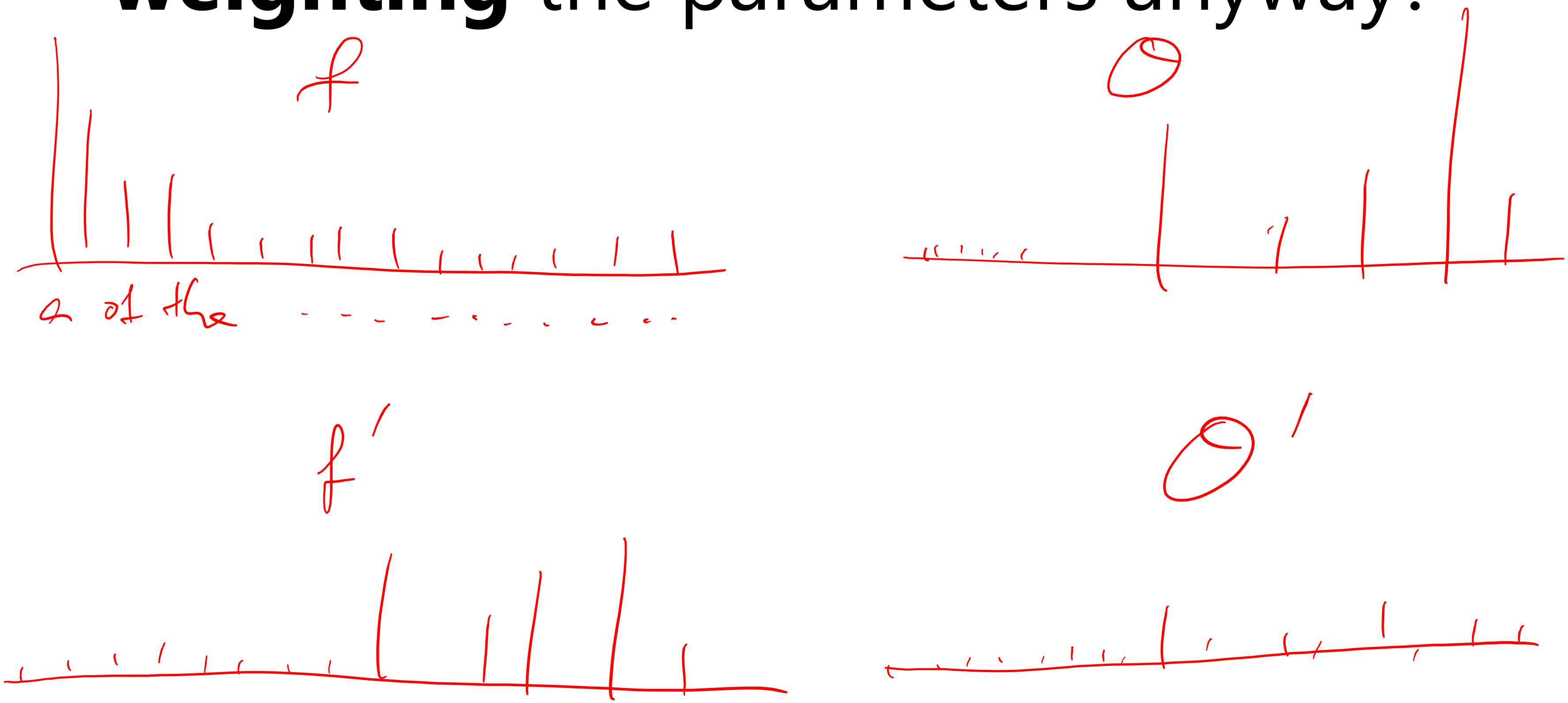
But what about when we're **weighting** the parameters anyway?

Yes!

- The **relative** weights of features is different between documents, so the two representations are not the same (up to scale)
- When we regularize, the scale of the features matters – if some "unimportant" features are very large, then the model can overfit on them "for free"

Finding relevant terms

But what about when we're **weighting** the parameters anyway?



Finding relevant terms

But what about when we're **weighting** the parameters anyway?

Questions?

Further reading:

- Original TF-IDF paper (from 1972)

"A Statistical Interpretation of Term Specificity and Its Application in Retrieval"

http://goo.gl/1CLwUV

CSE 158 – Lecture 11

Web Mining and Recommender Systems

Dimensionality-reduction approaches to document representation

Dimensionality reduction

How can we find **low-dimensional structure** in documents?

What we would like:

87 of 102 people found the following review helpful

☆☆☆☆☆ **You keep what you kill**, December 27, 2004

By Schtinky "Schtinky" (Washington State) - See all my reviews

VINE™ VOICE

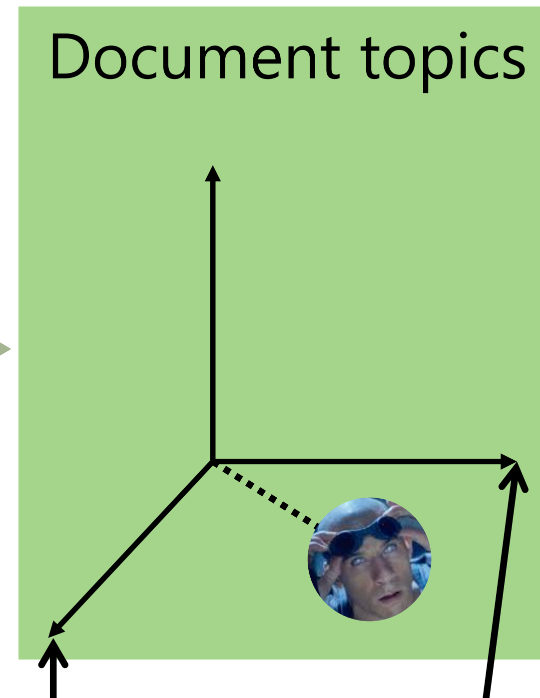
This review is from: The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from `Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to `Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

topic model



Action:
action, loud, fast, explosion,...

Sci-fi
space, future, planet,...

Singular-value decomposition

Recall (from lectures 3&5)

$$R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix}$$

(e.g.)
matrix of
ratings

(square roots of)
eigenvalues of RR^T

$$R = U\Sigma V^T$$

eigenvectors of RR^T

eigenvectors of R^TR

Singular-value decomposition

Taking the eigenvectors corresponding to the top-K eigenvalues is then the "best" rank-K approximation

$$R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix}$$

(square roots of top k) eigenvalues of RR^T

$$R \simeq U^{(k)} \Sigma^{(k)} V^{(k)T}$$

(top k) eigenvectors of RR^T

(top k) eigenvectors of $R^T R$

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

$$X = \left.\underbrace{\begin{pmatrix} 1 & 0 & \cdots & 4 \\ 0 & 2 & & 0 \\ 31 & 23 & & 97 \\ 0 & 98 & & 1 \\ 473 & 88 & & 347 \\ \vdots & & \ddots & \vdots \\ 11 & 34 & \cdots & 13 \end{pmatrix}}_{\text{documents}}\right\}\text{terms}$$

document matrix

X is a *TxD* matrix whose **columns** are bag-of-words representations of our documents

T = dictionary size
D = number of documents

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

$X^T X$ is a *D*x*D* matrix.

$U^{(k)}\Sigma^{(k)}$ is a low-rank approximation of each **document**

eigenvectors of $X^T X$

$X X^T$ is a *T*x*T* matrix.

$V^{(k)}\Sigma^{(k)}$ is a low-rank approximation of each **term**

eigenvectors of $X X^T$

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

$$R = \begin{bmatrix} \cdot & 3 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 5 & 2 & \cdot \\ \cdot & 1 & \cdot & 4 & \cdot \end{bmatrix} \sim \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

low-d user rep.

low-d item rep

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

$$D = \begin{bmatrix} 105 & & 1 & 2 \\ & & & 3 \\ 3 & 4 & & 612 \\ 1 & & & \end{bmatrix} = \gamma_w \cdot \gamma_d$$

words

docs

γ_w

low-d rep. of w

γ_d

low-d rep of doc

Singular-value decomposition

Using our low rank representation of each **document** we can...

- Compare two documents by their low dimensional representations (e.g. by cosine similarity)
- To retrieve a document (by first projecting the query into the low-dimensional document space)
- Cluster similar documents according to their low-dimensional representations
- Use the low-dimensional representation as features for some other prediction task

Singular-value decomposition

Using our low rank representation of each **word** we can...

- Identify potential synonyms – if two words have similar low-dimensional representations then they should have similar "roles" in documents and are potentially synonyms of each other
- This idea can even be applied across languages, where similar terms in different languages ought to have similar representations in parallel corpora of translated documents

Singular-value decomposition

This approach is called **latent semantic analysis**

- In practice, computing eigenvectors for matrices of the sizes in question is not practical – neither for XX^T nor X^TX (they won't even fit in memory!)
- Instead one needs to resort to some approximation of the SVD, e.g. a method based on stochastic gradient descent that never requires us to compute XX^T or X^TX directly (much as we did when approximating rating matrices with low-rank terms)

Probabilistic modeling of documents

Finally, can we represent documents in terms of the topics they describe?

What we would like:

87 of 102 people found the following review helpful

You keep what you kill, December 27, 2004

By Schtinky "Schtinky" (Washington State) - See all my reviews

VINE™ VOICE

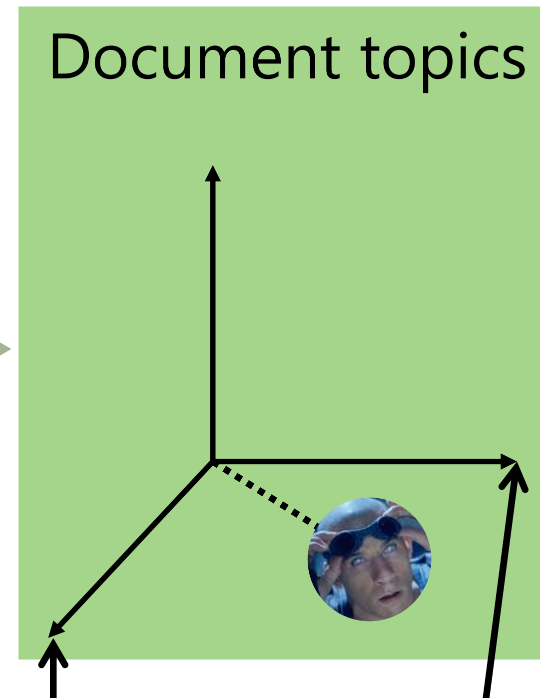
This review is from: The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from `Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to `Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

topic model



Action:
action, loud, fast, explosion,…

Sci-fi
space, future, planet,…

Probabilistic modeling of documents

Finally, can we represent documents in terms of the topics they describe?

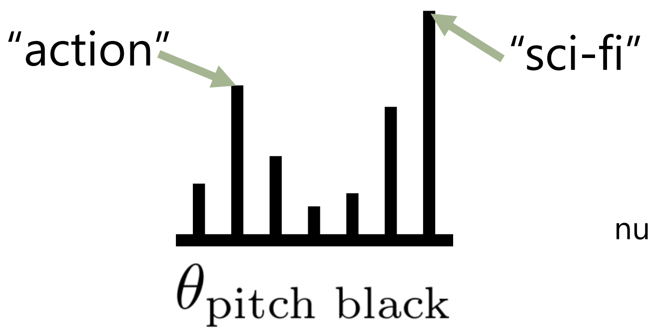
- We'd like each document to be a **mixture over topics** (e.g. if movies have topics like "action", "comedy", "sci-fi", and "romance", then reviews of action/sci-fis might have representations like [0.5, 0, 0.5, 0])

action sci-fi

- Next we'd like each topic to be a **mixture over words** (e.g. a topic like "action" would have high weights for words like "fast", "loud", "explosion" and low weights for words like "funny", "romance", and "family")

Latent Dirichlet Allocation

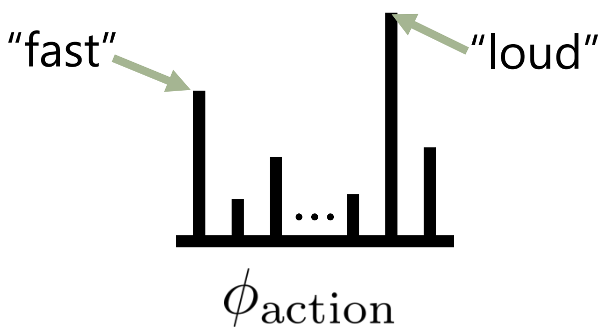
Both of these can be represented by **multinomial distributions**



Each document has a **topic distribution** which is a mixture over the topics it discusses

number of topics

$\theta_d \in \Delta^K$ i.e., $\forall_d \sum_k \theta_{d,k} = 1$



Each topic has a **word distribution** which is a mixture over the words it discusses

number of words

$\phi_k \in \Delta^D$ i.e., $\forall_k \sum_w \phi_{k,w} = 1$

Latent Dirichlet Allocation

Under this model, we can estimate the probability of a particular bag-of-words appearing with a particular topic and word distribution

document

iterate over word positions

probability of this word's topic

probability of observing this word in this topic

$$p(d|\theta, \phi, z) = \prod_{j=1}^{\text{length of } d} \theta_{z_{d,j}} \phi_{z_{d,j}, w_{d,j}}$$

Problem: we need to estimate all this stuff before we can compute this probability!

Latent Dirichlet Allocation

E.g. some topics discovered from an Associated Press corpus

labels are determined manually

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Latent Dirichlet Allocation

And the topics most likely to have generated each word in a document

labels are determined manually

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

From http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

Latent Dirichlet Allocation

Many many many extensions of Latent Dirichlet Allocation have been proposed:

- To handle temporally evolving data:

"Topics over time: a non-Markov continuous-time model of topical trends" (Wang & McCallum, 2006)
http://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf

- To handle **relational** data:

"Block-LDA: Jointly modeling entity-annotated text and entity-entity links" (Balasubramanyan & Cohen, 2011)
http://www.cs.cmu.edu/~wcohen/postscript/sdm-2011-sub.pdf

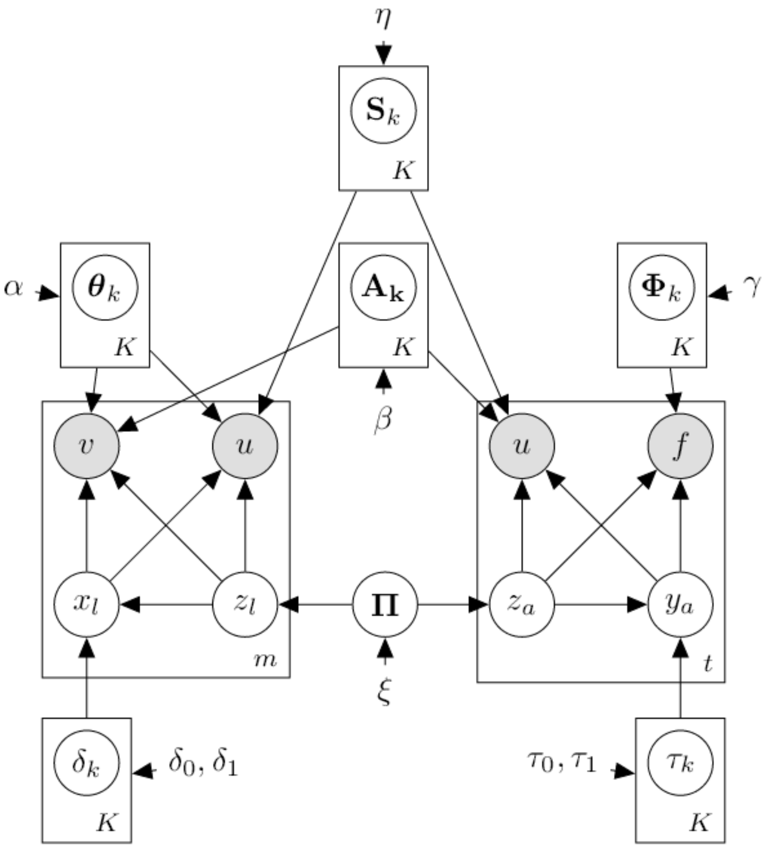
"Relational topic models for document networks" (Chang & Blei, 2009)
https://www.cs.princeton.edu/~blei/papers/ChangBlei2009.pdf

"Topic-link LDA: joint models of topic and author community" (Liu, Nicelescu-Mizil, & Gryc, 2009)
http://www.niculescu-mizil.org/papers/Link-LDA2.crc.pdf

Latent Dirichlet Allocation

Many many many extensions of Latent Dirichlet Allocation have been proposed:

"WTFW" model (Barbieri, Bonch, & Manco, 2014), a model for relational documents



Summary

Today…

Using **text** to solve predictive tasks

- Representing documents using bags-of-words and TF-IDF weighted vectors
- Stemming & stopwords
- Sentiment analysis and classification

Dimensionality reduction approaches:

- Latent Semantic Analysis

Questions?

Further reading:

- Latent semantic analysis

"An introduction to Latent Semantic Analysis" (Landauer, Foltz, & Laham, 1998)
http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

- LDA

"Latent Dirichlet Allocation" (Blei, Ng, & Jordan, 2003)
http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

- Plate notation

http://en.wikipedia.org/wiki/Plate_notation
"Operations for Learning with Graphical Models" (Buntine, 1994)
http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume2/buntine94a.pdf

A few assignment 1 tips

Task 1

~~Optimize MAE~~ / outliers

↓

overfitting

---

one-hot as a feature

A few assignment 1 tips

Task 2

- Men's / Women's
- features besides words